

Reviewer 2:

*"I carefully read the response and revised manuscript by Ruijsch et al. Unfortunately, I still cannot recommend publication, as several important issues remain unresolved. Overall, the manuscript mainly presents a comparison between LPJmL-5 simulations and satellite-derived datasets, but it remains unclear what new insights this comparison provides or what specific problem it resolves. More importantly, the analysis does not support the broader motivation outlined by the authors. The study aims to evaluate whether LPJmL-5 can be used to assess vegetation responses to multi-year droughts beyond the satellite observation period, potentially including longer historical periods or future projections. However, the current analysis is limited to comparisons within the relatively short satellite record, and therefore cannot demonstrate whether the model provides reliable estimates outside the observational period. As a result, it remains unclear how this comparison improves confidence in using the model to investigate vegetation responses to drought over longer timescales."*

We thank the reviewer for taking the time to review our manuscript. We agree that our analysis is limited to the satellite observation period and does not directly demonstrate LPJmL-5's reliability for historical periods outside of the satellite era or future projections. The primary aim of this study, however, is to benchmark LPJmL-5 against MODIS-derived GPP to systematically evaluate how well the model reproduces vegetation responses to multi-year and normal droughts. By having this information available we can begin to make estimates of LPJmL-5's ability for periods outside of the observational record, or at the very least have a quantification of the likely magnitude of the model error in those periods.

Secondly, this benchmarking provides insights into model strengths and weaknesses, including which plant functional types and regions are well represented and where the model tends to over- or under-estimate drought responses. It also highlights specific processes in the model, such as vegetation stress, recovery, and drought-induced mortality, that could be improved and thus provide valuable guidance for future model development.

By rephrasing the introduction to this study, we better emphasize that the comparison is not intended to validate long-term projections, but rather to establish a quantitative reference for interpreting modelled drought responses and to improve confidence in using the model within the observed climate period. We have revised the manuscript (Abstract, Introduction, and Conclusion) to make this framing clearer and to clarify the contribution of our work beyond a simple dataset comparison.

*"-abstract:*

*The description of cropland performance in the abstract is somewhat confusing. Earlier, the text states that model performance declines in regions with croplands, suggesting the worst agreement, while later it states that croplands show relatively good agreement. Although these statements appear to refer to different aspects of the analysis (absolute GPP versus drought-response dynamics), this distinction is not clearly explained in the abstract."*

We thank the reviewer for pointing out this source of confusion. In the revised abstract, we have clarified the distinction between absolute GPP performance and drought-response dynamics in croplands. Specifically, we now note that while LPJmL-5's reproduction of absolute GPP declines in cropland-dominated regions, the model captures drought-induced GPP anomalies in croplands relatively well. This wording makes explicit that the two statements refer to different aspects of model performance.

## *"-Introduction*

*Fourth paragraph: The rationale in this paragraph is not entirely convincing and should be better clarified. The text currently implies that DGVM simulations can compensate for the limited temporal coverage of satellite observations by extending vegetation records beyond the satellite era. However, model outputs cannot substitute observational records, and this framing is therefore conceptually problematic. As I understand it, the purpose here is to assess whether LPJmL can reproduce vegetation responses to multi-year drought events, using MODIS observations as a benchmark. If so, this motivation should be stated more clearly, emphasizing model evaluation against observations rather than presenting DGVMs as a replacement for short satellite records.*

*Fifth paragraph: However, the comparison with MODIS over a relatively short observational period cannot demonstrate whether LPJmL-5 outputs are reliable for longer historical periods or future projections. You need to compare more datasets with longer time period if for this purpose."*

We thank the reviewer for pointing this out. We agree that DGVM outputs cannot replace observational records, however they can provide information in periods when satellite information is not available. To clarify, the main purpose of our study is to evaluate how well LPJmL-5 reproduces observed vegetation responses to drought, using MODIS-derived GPP as a benchmark. We have revised the introduction and abstract to emphasize this benchmarking objective rather than implying that the model can substitute for short satellite records.

We also acknowledge that our comparison is limited to the satellite era and therefore cannot, on its own, demonstrate the reliability of LPJmL-5 for longer historical periods or future projections. While our study does not extend beyond the observational record, benchmarking LPJmL-5 in this way is a necessary first step for interpreting model simulation that cover periods outside of the observational record. Additionally, it helps in identifying processes that are currently not well captured in LPJmL-5 and thus require improvement, providing a foundation for future studies of vegetation responses to multi-year droughts.

## *"-Results*

*The evaluation presented relies mainly on correlations between MODIS GPP and other satellite products (GPP, EVI, and SIF). However, correlation alone is insufficient to demonstrate that the simulated or satellite-derived GPP reliably represents vegetation productivity. High correlations may simply reflect similar spatial patterns, while substantial differences may still exist in absolute values, variability, or amplitudes. Similarly, Figure 2 presents spatial patterns of aggregated performance metrics (KGE and its components), which provide a general overview of model performance but do not directly show how well the model reproduces the magnitude and temporal dynamics of GPP. In particular, these metrics summarize long-term statistics but do not reveal differences in actual values, interannual variability, or year-to-year fluctuations. A more comprehensive evaluation should include more direct comparisons of modeled and observed GPP, such as biases, RMSE, and other metrics that quantify differences in magnitude and variability.*

*Ln329-330: The conclusion that croplands show the best modelled drought response appears difficult to reconcile with the results shown in Figure 3, where croplands exhibit the weakest agreement in absolute GPP dynamics (lowest KGE). Since drought responses are derived from GPP anomalies, it is unclear how the model can reproduce drought responses most accurately for the vegetation type where the overall GPP dynamics are least well simulated. In addition, Figure 8 mainly presents the distribution of response differences, but does not provide quantitative metrics (e.g., bias or RMSE) to support the claim that croplands perform best. Further clarification and more robust evaluation would be needed to support this conclusion. In addition, Figure 8 mainly presents the distribution of response differences, but does not provide quantitative metrics (e.g., bias or RMSE) to support the*

*claim that croplands perform best. Further clarification and more robust evaluation would be needed to support this conclusion."*

We thank the reviewer for highlighting the importance of evaluating both magnitude and variability in GPP. We would like to clarify that our manuscript already addresses a diverse range of metrics to capture different dynamic performance and covers the following aspects:

- Figure 2 presents the Kling-Gupta Efficiency (KGE) and its components (correlation, bias, and variability), which capture both the magnitude and variability of modelled GPP relative to MODIS.
- Figure 3 shows the cumulative distribution function of the KGE, providing insight into model performance across all grid cells.
- Figure 4 presents monthly GPP time series across six focus regions, demonstrating interannual variability, seasonal peaks, and drought responses. Correlations are computed temporally, ensuring that they reflect year-to-year dynamics rather than spatial patterns.
- Figures 6-8 include timeseries, spatial maps, and KDE plots of drought-induced GPP anomalies for different Plant Functional Types (PFTs), illustrating magnitude, temporal correlation, spatial correlation, and variability of drought responses.

A summary of the metrics used in each figure is provided in the manuscript for clarity:

|        |                   |   |
|--------|-------------------|---|
| Fig. 2 | GPP               | KGE metric<br>Correlation<br>Bias<br>Variability                      |
| Fig. 3 | GPP               | Cumulative distribution function                                      |
| Fig. 4 | GPP               | Temporal correlation<br>Absolute magnitude<br>Interannual variability |
| Fig. 5 | GPP <sub>SA</sub> | Temporal drought response timing                                      |
| Fig. 6 | GPP <sub>SA</sub> | Absolute magnitude and temporal correlation of MYD response           |
| Fig. 7 | GPP <sub>SA</sub> | Distribution, variability and spatial correlation of MYD response     |
| Fig. 8 | GPP <sub>SA</sub> | RMSE and distribution of MYD response                                 |

Regarding croplands, we acknowledge the potential confusion between absolute GPP performance and anomaly-based drought responses. To address this:

- We have clarified in the Discussion (lines 366–370) that LPJmL-5 reproduces drought-induced anomalies in croplands comparatively well despite weaker agreement in absolute GPP, indicating that biases in absolute productivity do not necessarily translate to errors in anomaly-based drought responses.
- In the revised Figure 8, we now include RMSE metrics for each land cover type, which quantitatively confirm that croplands exhibit the lowest RMSE in drought-response anomalies.

